

STORAGE CONTROLLER, DATA STORAGE SYSTEM CONTAINING IT AND DOUBLE-PAIR SUPPRESSION METHOD

Publication number: JP8305500

Publication date: 1996-11-22

Inventor: JIEIMUZU RINKAAN ISUKIYAN; ROBAATO
FUREDERITSUKU KEEN; UIRIAMU FURANKU
MITSUKA; ROBAATO UEZURII SHIYOMURAA

Applicant: IBM

Classification:


- international: G06F12/16; G06F3/06; G06F11/20; G06F12/08;
G06F12/16; G06F3/06; G06F11/20; G06F12/08; (IPC1-
7): G06F3/06; G06F12/08; G06F12/16

- european: G06F11/20L4M10

Application number: JP19960088562 19960410

Priority number(s): US19950424930 19950419

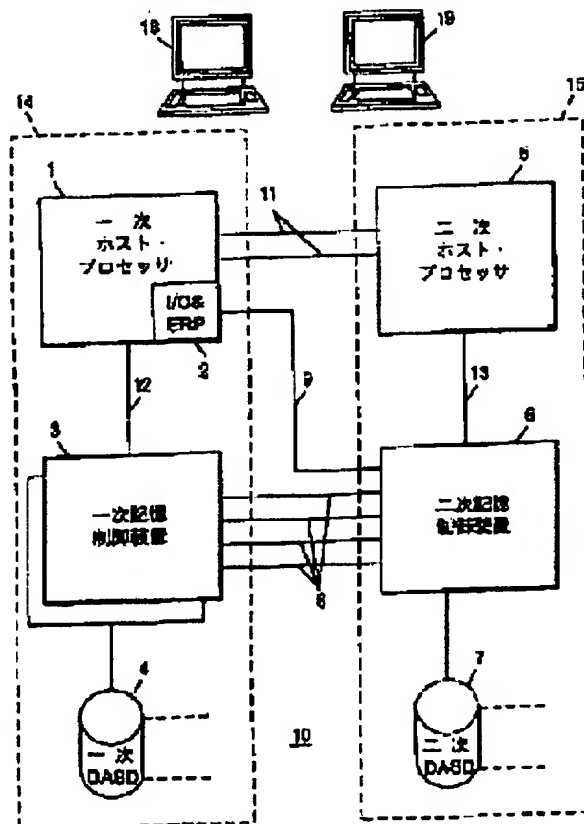
Also published as:

 US5692155 (A1)

[Report a data error here](#)

Abstract of JP8305500

PROBLEM TO BE SOLVED: To provide a data storage system for suppressing plural double pairs across a single or plural storage sub-systems like an atomic. **SOLUTION:** A double pair is suppressed so that data can be maintained on their secondary DASD 7 in a sequence matching sequence. A host processor generates a record to be written in a primary DASD 4 of the double pair and record update. A storage controller instructs the copy of the record and the record update to the secondary DASD 7 of the double pair. The sequence compatibility can be maintained in the secondary DASD 7 by suppressing the double pair. During the suppression of the double pair, progressing writing I/O is also completed to the primary DASD. The storage controller generates a long busy signal for the following writing request for rejecting the following writing I/O from the host processor. During the suppression of the double pair due to the change of the recording, an application instructs the storage controller to mark the physical address of the primary DASD to be updated between the suppressing time and a resetting time.



Data supplied from the *esp@cenet* database - Worldwide

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平8-305500

(43) 公開日 平成8年(1996)11月22日

(51) Int.Cl. ⁶	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 3/06	3 0 4		G 0 6 F 3/06	3 0 4 F 3 0 4 M
12/08	3 1 0	7623-5B	12/08	3 1 0 Z
12/16	3 1 0	7623-5B	12/16	3 1 0 M

審査請求 未請求 請求項の数15 OL (全 20 頁)

(21) 出願番号 特願平8-88562

(22) 出願日 平成8年(1996)4月10日

(31) 優先権主張番号 4 2 4 9 3 0

(32) 優先日 1995年4月19日

(33) 優先権主張国 米国 (U S)

(71) 出願人 390009531

インターナショナル・ビジネス・マシーンズ・コーポレーション

INTERNATIONAL BUSINESS MACHINES CORPORATION

アメリカ合衆国10504、ニューヨーク州
アーモンク (番地なし)

(72) 発明者 ジェイムズ・リンカーン・イスキャン

アメリカ合衆国アリゾナ州、ツーソン、エヌ・ストーンハウス・プレイス 5190

(74) 代理人 弁理士 合田 潔 (外2名)

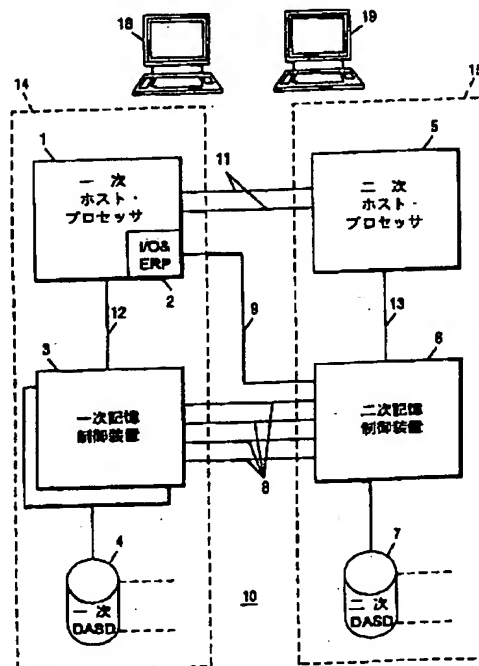
最終頁に続く

(54) 【発明の名称】 記憶制御装置、それを含むデータ記憶システムおよび二重ベア抑止方法

(57) 【要約】 (修正有)

【課題】単一又は多数の記憶サブシステムに跨る多数の二重ベアをアトミックに抑止するデータ記憶システムを提供する。

【解決手段】二重ベアは、それらの二次DASD上にデータがシーケンス整合した順序で維持されるように抑止される。ホスト・プロセッサは二重ベアの一次DASDに書き込まれるべきレコード及びレコード更新を発生する。記憶制御装置は二重ベアの二次DASDにレコード及びレコード更新の複写を指示する。シーケンス整合性は二重ベアを抑止することにより二次DASDにおいて維持される。二重ベアの静止は、進行中の書込みI/Oも一次DASDに対して完了する。記憶制御装置は、その後の書込みリクエストに対して長いビジー信号を生じることによってホスト・プロセッサからのその後の書込みI/Oも閉め出す。レコーディングの変更による二重ベアの抑止は、アプリケーションが、抑止する時間と再設定する時間との間に更新する一次DASDの物理的アドレスをマークするよう記憶制御装置に指示する。



(2)

特開平8-305500

1

【特許請求の範囲】

【請求項1】レコード及びレコード更新を書き込むこと及びバックアップの目的で前記レコード及びレコード更新を複写することができるデータ記憶システムにおいて、前記データ記憶システムはレコード及びレコード更新を発生するアプリケーションを走らせるホスト・プロセッサと、一次データ記憶装置及び二次データ記憶装置を有する第1二重ベアと、一次データ記憶装置及び二次データ記憶装置を有する第2二重ベアと、前記ホスト・プロセッサと前記第1二重ベア及び第2二重ベアの前記一次データ記憶装置との間に接続された記憶制御装置とを含み、レコード及びレコード更新の複写が進行中である時に第1及び第2二重ベアを抑止するための方法にして、

前記アプリケーションから前記第1二重ベアの一次データ記憶装置への将来のレコード及びレコード更新の書き込みを禁止するために前記ホスト・プロセッサからの初期静止コマンドに応答して前記記憶制御装置によって前記第1二重ベアを静止させるステップと、

前記アプリケーションから前記第2二重ベアの一次データ記憶装置への将来のレコード及びレコード更新の書き込みを禁止するために及びシーケンス整合した順序における前記第1及び第2二重ベアの前記二次データ記憶装置に複写されたレコード及びレコード更新を同期化するために前記ホスト・プロセッサからのその後の静止コマンドに応答して前記記憶制御装置によって前記第1二重ベアを静止させるステップと、

前記アプリケーションから前記第1及び第2二重ベアの一次データ記憶装置に送られたレコード及びレコード更新を前記第1及び第2二重ベアの二次データ記憶装置に前記記憶制御装置によって複写することを禁止するために、前記記憶制御装置が前記ホスト・プロセッサから抑止コマンドを受けることによって前記第1及び第2二重ベアを抑止するステップと、

前記第1及び第2二重ベアが抑止されること及びその後のレコード及びレコード更新が前記第1及び第2二重ベアの前記二次データ記憶装置に前記レコードを複写することなく前記第1及び第2二重ベアを前記一次データ記憶装置に書き込み可能であることを前記記憶制御装置によって前記アプリケーションに信号するステップと、を含む方法。

【請求項2】前記静止させるステップは前記アプリケーションから前記二重ベアの一次データ記憶装置へのレコード及びレコード更新のその後の書き込みを禁止するための長いビジー信号を前記記憶制御装置から前記ホスト・プロセッサに発生するステップを含むことを特徴とする請求項1に記載の方法。

【請求項3】前記記憶制御装置は、前記二重ベアが再設定される場合、前記ホスト・プロセッサから前記一次データ記憶装置に転送されたその後のレコード及びレコー

2

ド更新が前記二次データ記憶装置へのその後の複写のためにマークされないように前記二重ベアを終了させることを特徴とする請求項1に記載の方法。

【請求項4】前記記憶制御装置は、前記二重ベアが再設定される場合、前記レコード及びレコード更新が前記二次データ記憶装置に複写されるように、前記二重ベアの抑止の後に前記ホスト・プロセッサから前記一次データ記憶装置に転送されたレコード及びレコード更新をマークすることを特徴とする請求項1に記載の方法。

10 【請求項5】レコード及びレコード更新を第1及び第2二重ベアに複写することができるデータ処理システムにおいてレコード及びレコード更新の複写が進行中である時に各二重ベアを抑止するための記憶制御装置にして、前記データ記憶システムはアプリケーションを走らせるホスト・プロセッサを含み、前記第1及び第2二重ベアの各々は一次データ記憶装置及び二次データ記憶装置を有し、前記記憶制御装置はホスト・プロセッサと前記第1及び第2二重ベアの前記一次データ記憶装置との間に接続され、前記アプリケーションはレコード及びレコード更新を発生し及び静止二重ベア・コマンド及び抑止二重ベア・コマンドを発生し、前記一次データ記憶装置は前記レコード及びレコード更新を記憶し、前記二次データ記憶装置は前記レコード及びレコード更新の複写を記憶するものにおいて、

20 前記ホスト・プロセッサと前記第1及び第2二重ベアとの間のレコード及びレコード更新を指示するための記憶装置バスであって、前記アプリケーションからの前記静止二重ベア・コマンドに応答して前記第1及び第2二重ベアを静止させ、前記レコード及びレコード更新が前記二重ベアの各々における前記二次データ記憶装置の各々に複写され且つシーケンス整合順序で同期化されるように前記アプリケーションからの前記抑止二重ベア・コマンドに応答して前記第1及び第2二重ベアを抑止する記憶装置バスと、

前記記憶装置バスに接続され、前記二重ベアの二次データ記憶装置に複写されるべき前記レコード及びレコード更新を記憶するためのメモリと、を含む記憶制御装置。

40 【請求項6】前記メモリはキャッシュ・メモリであることを特徴とする請求項5に記載の記憶制御装置。

【請求項7】レコード及びレコード更新をシーケンス整合した順序で記憶するための不揮発性記憶装置(NV S)を含み、前記レコード及びレコード更新はその後前記二重ベアの二次データ記憶装置に複写されることを特徴とする請求項5に記載の記憶制御装置。

【請求項8】前記二重ベアが抑止される時、前記レコード更新を受ける前記二重ベアと関連した物理的アドレスをマークするために前記NV Sにおけるビットマップを含み、前記レコード更新は前記二重ベアが抑止されている時に前記一次データ記憶装置に書き込まれ、一旦前記

(3)

特開平8-305500

3

二重ペアが再設定されると前記二次データ記憶装置に複写されることを特徴とする請求項7に記載の記憶制御装置。

【請求項9】前記記憶装置バスは、前記ホスト・プロセッサから前記二重ペアの一次データ記憶装置に現在書き込まれている前記レコード更新が完了することを可能にするために、及びその後のレコード更新が前記ホスト・プロセッサから前記二重ペアの一次データ記憶装置に書き込まれることを禁止するために、前記静止二重ペア・コマンドに応答して長いビジー信号を前記ホスト・プロセッサに発生することを特徴とする請求項5に記載の記憶制御装置。

【請求項10】複数のレコード更新を発生するアプリケーションを走らせ、及び静止二重ペア・コマンド及び抑止二重ペア・コマンドを発生するホスト・プロセッサと、

前記ホスト・プロセッサに接続されたチャネルと、
各々が一次データ記憶装置及び二次データ記憶装置を有する第1及び第2二重ペアであって、前記一次データ記憶装置の各々は複数のレコード更新を記憶し、前記二次データ記憶装置の各々は前記複数のレコード更新の複写を記憶するものと、

前記チャネルによって前記ホスト・プロセッサに接続され、更に、前記第1及び第2二重ペアの一次データ記憶装置の各々に接続された記憶制御装置と、

を含み、前記記憶制御装置は、

前記ホスト・プロセッサから転送された複数のレコード更新を最初に記憶するためのメモリと、

前記チャネルと前記一次データ記憶装置の各々との間に接続された記憶装置バスと、

を含み、

前記メモリは前記記憶装置バスに接続されること、及び前記記憶装置バスは前記ホスト・プロセッサと前記一次データ記憶装置の間で前記メモリを介して前記複数のレコード更新の移動を指示し、前記ホスト・プロセッサから前記静止二重ペア・コマンド及び前記抑止二重ペア・コマンドを受け取り、前記複数のレコード更新をシーケンス整合した順序で同期化するために前記複数のレコード更新の複写を前記二次データ記憶装置に転送すること、
を特徴とするデータ記憶システム。

【請求項11】前記記憶制御装置におけるメモリはキャッシュ・メモリであることを特徴とする請求項10に記載のデータ記憶システム。

【請求項12】前記記憶制御装置において前記記憶装置バスに接続された不揮発性記憶装置(NVS)を含み、前記NVSはレコード及びレコード更新をシーケンス整合した順序で記憶すること及び前記レコード及びレコード更新は前記二重ペアの二次データ記憶装置に複写されることを特徴とする請求項10に記載のデータ記憶シ

4

テム。

【請求項13】前記二重ペアの各々の各二次データ記憶装置は前記一次データ記憶装置の各々に対して遠隔の二次サイトに置かれること、及び前記二次サイトは前記一次サイトにおける前記記憶制御装置から前記レコード及びレコード更新を受けるための及び前記各二次データ記憶装置に接続された記憶制御装置を含むことを特徴とする請求項10に記載のデータ記憶システム。

10 【請求項14】前記二重ペアの各々における前記各二次データ記憶装置は前記一次データ記憶装置の各々に対して遠隔の二次サイトに置かれること、及び前記二次サイトはホスト・プロセッサと、

前記ホスト・プロセッサに接続されたチャネルと、

前記チャネルによって前記ホスト・プロセッサに接続され、更に前記各二次データ記憶装置に接続された記憶制御装置と、

を含むことを特徴とする請求項10に記載のデータ記憶システム。

20 【請求項15】前記記憶制御装置における前記記憶装置バスは、前記ホスト・プロセッサから前記二重ペアの一次データ記憶装置に現在書き込まれている前記レコード更新が完了することを可能にするために、及びその後のレコード更新が前記ホスト・プロセッサから前記二重ペアの一次データ記憶装置に書き込まれることを禁止するために、前記静止二重ペア・コマンドに応答して長いビジー信号を前記ホスト・プロセッサに発生することを特徴とする請求項10に記載のデータ記憶システム。

【発明の詳細な説明】

【0001】

30 【発明の属する技術分野】本発明は、概して云えば、データ記憶技法に関するものであり、更に詳しく云えば、多数の装置又は装置サブシステムにまたがる二重複写オペレーション或いは遠隔二重複写オペレーションをアトミックに抑止するためのシステムに関するものである。

【0002】

40 【従来の技術】データ処理システムは、処理データと関連して、一般に、大量のデータを記憶する必要がある。そのデータは、効果的にアクセスされ、修正され、そして再記憶され得るものである。データ記憶装置は、一般に、効率的且つコスト効果的なデータ記憶装置を提供するために、幾つかの異なるレベルに或いは階層的に分割される。第1レベルの又は最高レベルのデータ記憶装置は、通常、動的又は静的ランダム・アクセス・メモリ(DRAM又はSRAM)と呼ばれる電子的メモリを包含する。電子的メモリは、数百万バイトのデータが各回路上に記憶可能であり且つそのようなデータ・バイトへのアクセスがナノ秒単位で測定されるという半導体集積回路の形式を取る。その電子的メモリは、アクセスが全体的に電子的であるので、データに対する最速のアクセスを与える。

(4)

特開平8-305500

5

6

【0003】第2レベルのデータ記憶装置は、通常、直接アクセス記憶装置(DASD)を包含する。DASDは、例えば、データのビットを形成する「1」又は「0」を表すためのディスク表面におけるミクロン単位の大きさの磁氣的に又は光学的に変更されたスポットとしてデータのビットを記憶する磁氣的又は光学的ディスクより成る。磁氣的DASDは、残留磁気材料でもって被覆された1つ又は複数個のディスクを有する。それらのディスクは、保護された環境内で回転可能に装着される。各ディスクは、多くの同心円トラック、即ち、密接した間隔の円に分けられる。データは、各トラックに沿ってビット毎に順次に記憶される。ハード・ディスク・アセンブリ(HDA)として知られたアクセス機構は、一般に、1つ又は複数個の読取り/書込みヘッドを有し、ディスクがそれらの読取り/書込みヘッドを通して回転する時、トラックを横切って移動してそれらのディスクの表面との間でデータを転送するように各DASDに設けられる。DASDは数ギガバイトのデータを記憶することができ、そのようなデータへのアクセスは一般にミリ秒単位(電子的メモリよりも数桁の大きさに遅い)で測定される。DASDに記憶されたデータへのアクセスは、ディスク及びHDAを所望のデータ記憶ロケーションに物理的に位置づける必要があるために遅くなる。

【0004】第3レベル、即ち低レベルのデータ記憶装置はテープ・ライブラリ、又はテープ及びDASDライブラリを含む。データに対するアクセスはライブラリではずっと遅くなる。それは、ロボットがその必要なデータ記憶媒体を選択し及びロードする必要があるためである。利点は、非常に大きいデータ記憶容量、例えば、テラバイトのデータ記憶のためにコストが減少することである。テープ記憶装置はバックアップ目的で使用されることが多い。即ち、第2レベルの階層に記憶されたデータは磁気テープにおいて保護するために複写される。テープ又はライブラリに記憶されたデータへのアクセスは、現在、秒単位のレベルである。

【0005】バックアップ・データを複写させることは、多くのビジネスにとって必須のことである。それは、ビジネスにとってデータ喪失が大災害になり得るためである。一次記憶レベルで失われたデータを回復するために必要な時間も重要な回復の考察事項である。テープ又はライブラリ・バックアップにおける速度の改良は二重複写を含んでいる。二重複写の一例は追加のDASDを設けて、データがその追加のDASDに書き込まれるようにする(ミラーリングとも呼ばれる)。そこで、一次DASDが故障すると、二次DASDがデータに関して頼みにされる。この方法の欠点は、必要なDASDの数が2倍になることである。

【0006】2倍の記憶装置を設ける必要性を克服するもう1つのデータ・バックアップ方法は、低価格ディス

クの冗長配列(RAID)構成またはディスク・アレイにデータを書き込むことに関連する。この場合、データは、それが多くのDASDの間で配分されるように書き込まれる。1つのDASDだけが障害を生じた場合、失われたデータは残りのデータ及びエラー訂正手順を使用することによって回復可能である。現在、幾つかの異なるRAID構成が利用可能である。

【0007】前述のバックアップ解決法は、一般には、記憶装置又は媒体が障害を生じた場合にデータを回復させるには十分である。これらのバックアップ方法は装置の障害に対しては有用である。それは、二次データが一次データのミラーであるためである。即ち、二次データは一次データと同じボリューム通し番号(VOLSER)を有する。一方、システム障害回復は、ミラーされた二次データを使用して得ることはできない。従って、システム全体或いはその場所を破壊するような災害、例えば、地震、火事、爆発、台風等が発生した場合、データを回復するためには更なる保護が必要である。災害回復は、データの二次複写が一次データから離れたロケーションに保管されることを必要とする。災害保護を行う既知の方法は、日毎又は週毎をベースにデータをテープにバックアップすることである。そこで、そのテープは搬送手段によって取り上げられ、通常、一次データ・ロケーションから数キロメートル離れた安全な保管領域に持って行かれる。このバックアップ方法では、バックアップ・データを検索するのに数日を必要とするという問題、及び数時間又は数日分のデータが失われ或いは不良化している間に、保管ロケーションが同じ災害によって破壊されることがあるという問題が提起されている。わずかに改良されたバックアップ法は、毎晩、データをバックアップ・ロケーションに搬送するものである。これは、更に離れたロケーションにデータを記憶することを可能にする。バックアップが二重複写法におけるように継続的には生じないため、バックアップとバックアップとの間に或データが失われることがある。従って、ユーザにとっては許容し得ないかなりの量のデータが失われることがある。

【0008】更に最近紹介されたデータ災害回復法は遠隔二重複写を含む。その方法では、データは遠く離れてバックアップされるのみならず、連続的に(同期的に、或いは非同期的に)もバックアップされる。1つのホスト・プロセッサから他のホスト・プロセッサに、或いは、1つの記憶制御装置から他の記憶制御装置に、或いはそれらの組合せでその二重化されたデータをコミュニケーションするためには、そのプロセスを実現するためのかなりの量の制御データが必要である。しかし、高いオーバーヘッドは一次サイトの処理を維持するために二次サイトを損なうことがあり、従って、災害が発生した場合、二次サイトが一次サイトを回復させることができるという能力に影響することがある。

(5)

特開平8-305500

7

【0009】二重複写は、DASDサブシステムにおいて二重ベアを形成する一次ボリューム及び二次ボリュームを必然的に伴う。DASDサブシステムにおける複写動作は、その複写されたボリュームに対するI/Oコマンドによって制御される。そのようなI/Oコマンドは、二重ベアを設定又は抑止するために、或いは二重ベアのステータスを待ち行列化するために装置毎の制御を行う。しかし、装置毎の制御がすべての災害回復アプリケーションにとって十分であるわけではない。二次ロケーションにおける複写されたデータは、その複写されたデータが元のデータと時間的整合（順次整合）している限り使用可能である。歴史的には、I/Oコマンドはデータを複写する間そのシステムを停止させることによってそのような整合性を確保し、従って、データに対する更なる更新が生じなかったことを保証した。この方法に対する改良はT0又は同時複写として知られ、それはシステムが停止している時間を減少させるが、依然として抑止が必要であった。

【0010】実時間二重複写、例えば、拡張遠隔複写（XRC）又は同格（ピアツーピア）通信遠隔複写（PPRC）は、それらの複写が災害回復にとって使用可能であるように二次ボリュームにまたがる時間的整合性を保証する。しかし、もう一度云うが、一次システムの抑止は装置毎の制御にとって必要なことである。これらの抑止は或システムにおいては受け入れ難い崩壊を生じさせる。XRCシステムは、ソフトウェア制御のデータ・ムーバを介した解決法を与える。その場合、二次装置が部分的に複写オペレーションの非同期的性質によって時間整合するように、単一のコマンドがセッションを停止させる。このデータ・ムーバの解決法は、データをボリューム毎に同期的に複写する二重複写又はPPRCバックアップ環境においては利用し得ない。各ボリュームは如何なる他のボリュームからも独立した複写であるが、一組のボリュームにおけるデータが論理的に従属することはあり得る。

【0011】従って、必要なことは、一組の独立した装置に対する複写動作のシーケンス整合の抑止を、システム・オペレーション抑止を必要とすることなく、システム・オペレーションにおける最小の遅延でもって生じさせることができる災害回復システムである。

【0012】

【発明が解決しようとする課題】本発明の目的は、同期的二重複写記憶システムにおいて多数の装置をアトミックに制御するための改良された設計及び方法を提供することにある。

【0013】本発明のもう1つの目的は、二次データ記憶装置におけるデータの論理的従属性を維持しながら、システム・オペレーションを抑止することなく多数の同期的二重ベアを静止させるための改良された方法及びシステムを提供することにある。

8

【0014】

【課題を解決するための手段】本発明の第1実施例によれば、レコード及びレコード更新の複写が進行中である間、第1及び第2二重ベアを抑止するための方法が与えられる。その方法を実行するためのデータ記憶システムは、レコード及びレコード更新を生じさせるアプリケーションを実行するホスト・プロセッサを含む。第1二重ベアは一次データ記憶装置及び二次データ記憶装置を含み、一方、第2二重ベアも一次データ記憶装置及び二次データ記憶装置を含む。ホスト・プロセッサと第1及び第2二重ベアの一次データ記憶装置との間には記憶制御装置が結合される。その方法は、記憶制御装置においてホスト・プロセッサからの初期静止コマンドにตอบสนองしてアプリケーションから第1二重ベアの一次データ記憶装置へのその後のレコード及びレコード更新の書込みを禁止することによって第1二重ベアを静止させる。

【0015】又、その方法は、記憶制御装置においてホスト・プロセッサからのその後の静止コマンドにตอบสนองしてアプリケーションから第2二重ベアの一次データ記憶装置へのその後のレコード及びレコード更新の書込みを禁止することによって第2二重ベアを静止させる。これは、シーケンス整合した順序で第1及び第2二重ベアの二次データ記憶装置に複写されたレコード及びレコード更新を調整する。そこで、第1及び第2二重ベアはホスト・プロセッサから抑止コマンドを受けた後に記憶制御装置によって抑止され、複写動作は禁止される。そこで、記憶制御装置は、第1及び第2二重ベアが抑止されそして二次データ記憶装置にレコードを複写することなくその後のレコード及びレコード更新が一次データ記憶装置に書き込まれる。しかし、二重ベアは最初に静止させられたので、すべての二次データ記憶装置が特定の時点においてアプリケーションに関してシーケンス整合する。

【0016】本発明のもう1つの実施例では、レコード及びレコード更新を第1及び第2二重ベアに複写するための、及びレコード及びレコード更新の複写が進行中である間、各二重ベアを抑止するための記憶制御装置がデータ記憶システムに設けられる。データ記憶システムは、アプリケーションを実行するホスト・プロセッサを含み、各二重ベアは、更に一次データ記憶装置及び二次データ記憶装置を含む。記憶制御装置はホスト・プロセッサと各一次データ記憶装置との間に結合される。アプリケーションはレコード及びレコード更新を発生し、静止二重ベア・コマンド及び抑止二重ベア・コマンドを発生する。レコード及びレコード更新は一次記憶装置に記憶され、一方、それらの各複写は二次記憶装置に記憶される。

【0017】記憶制御装置における記憶装置バスは、ホスト・プロセッサと第1及び第2二重ベアとの間のレコード及びレコード更新を指示する。更に、記憶装置バス

(6)

特開平8-305500

9

は、アプリケーションからの静止二重ペア・コマンドに
応答して第1及び第2二重ペアを静止させる。レコード
及びレコード更新が各二次データ記憶装置に複写され且
つシーケンス整合した順序で同期化されるように、ア
プリケーションからの抑止二重ペア・コマンドに
第1及び第2二重ペアが記憶装置バスによって抑止され
る。二重ペアの前記二次データ記憶装置に複写されるべ
きレコード及びレコード更新を最初に記憶するためのメ
モリが設けられる。

【0018】

【発明の実施の形態】代表的なデータ処理システムは、
データを計算及び操作するための、例えば、少なくとも
1つのIBM3990記憶制御機構を接続されたデータ
機能記憶管理サブシステム/多重仮想システム(DFS
MS/MVS)ソフトウェアを実行するためのIBMシ
ステム/370或いはIBMシステム/390のような
ホスト・プロセッサの形式を取るものでよい。一般に、
記憶制御装置はメモリ・コントローラ及びそれに組み込
まれた1つ又は複数のキャッシュ・メモリ・タイプを含
む。記憶制御装置は、更に、IBM3380又は33
90のような直接アクセス記憶装置(DASD)のグル
ープに接続される。ホスト・プロセッサは大きな計算力
を与えるけれども、記憶制御装置は効率的に転送するた
めに、ステージ/デステージするために、変換するた
めに、及び一般に大きなデータ・ベースをアクセスする
ために必要な機能を与える。

【0019】データ記憶システムにおいてデータを保護
する1つの形式は1つのDASDのデータを他のDAS
D上にバックアップすることである。データ・バックア
ップを行うためのこの手順は、一般に、二重複写と呼ば
れる。データは一次DASDに記憶され、二次DASD
に複写される。それらの一次及び二次DASDは二重ペ
アを形成する。二重複写二重ペアの一次及び二次装置は
別個の装置ストリングにグループ分け可能であるが、そ
れらは同じ記憶制御装置に接続されなければならない。
その同じ記憶制御装置に接続される要件のために、二重
複写は、データ・バックアップのための受け入れ可能な
方法であるが、それは災害回復システムにとって実施可
能ではない。

【0020】一般的なデータ処理システムに対する災害
回復保護は、一次DASDに記憶された一次データが二
次又は遠隔ロケーションにバックアップされることを必
要とする。一次及び二次ロケーションを分ける距離はユ
ーザにとって受け入れ可能なリスクのレベルに依存し、
数キロメートルから数千キロメートルまで変わり得るも
のである。二次又は遠隔ロケーションは、バックアップ
・データの複写を行うことに加えて、一次システムがデ
ィスエーブルされた場合に一次システムに対する処理を
引き継ぐのに十分なシステム情報も持たなければなら
ない。これは、単一の記憶制御装置が一次及び二次サイト

10

において一次及び二次DASDストリングの両方にはデ
ータを書き込まないために、部分的には当然である。代
わりに、一次データが一次記憶制御装置に接続された一
次DASDストリング上に記憶され、一方、二次データ
が二次記憶制御装置に接続された二次DASDストリン
グ上に記憶される。

【0021】二次サイトは一次サイトから十分に離れて
いなければならないのみならず、一次データを実時間で
バックアップできなければならない。一次データが或最
小の遅延で更新される時、二次サイトは一次データをバ
ックアップする必要がある。更に、二次サイトは、一次
サイトで走っていてデータ又は更新を発生するアプリケ
ーション・プログラム(例えば、IMS、DB2)に関
係なく、一次データをバックアップしなければならない。
二次サイトで必要とされる困難なタスクは、二次デー
タが順序整合していなければならないことである。即
ち、二次データは一次データと同じ順次順序(順次整
合)で複写される。それはかなりのシステム考察を必要
とする。順次整合性は、多数の記憶制御装置(その各々
がデータ処理システムにおける多数のDASDを制御す
る)の存在によって複雑にされる。順次整合がない場
合、一次データと整合しない二次データがその結果とし
て生じ、従って災害回復を不良化するであろう。

【0022】遠隔のデータ二重化は2つの一般的なカテ
ゴリ、即ち、同期的なもの及び非同期的なものに分けら
れる。同期的遠隔複写は、一次データを二次ロケーシ
ョンに送ること、及び一次DASD入出力(I/O)オペ
レーションを終了する(一次ホストにチャネル終了(C
E)及び装置終了(DE)を与える)前にそのようなデ
ータの受信を確認することを含む。従って、二次の確認
を待っている間、同期的複写は一次DASDのI/O応
答時間を遅らせる。一次I/O応答遅延は一次システム
と二次システムとの間の距離に比例して増加する(遠隔
距離を数十キロメートルに制限する要因)。しかし、同
期的複写は、比較的わずかなシステム・オーバーヘッドで
もって二次サイトにおいて整合したデータを順次与える
。

【0023】非同期的遠隔複写は、データが二次サイト
において確認される前に一次DASDのI/Oオペレー
ションが完了する(一次ホストにチャネル終了(CE)
及び装置終了(DE)を与える)ので、一次アプリケ
ーション・システムのより良いパフォーマンスを与える。
従って、一次DASDのI/O応答時間は二次サイトま
での距離に依存せず、二次サイトは一次サイトから数千
キロメートルも離れていてもよい。しかし、二次サイト
で受信されるデータは一次更新の順序ではないことがよ
くあるので、データ・シーケンスの整合性を確保する
ためには大量のシステム・オーバーヘッドが必要である。
一次サイトにおける障害は、一次ロケーションと二次ロ
ケーションとの間で伝送中であつたデータが失われると

(7)

特開平8-305500

11

いう結果を生じることがある。

【0024】(a) 非同期的遠隔複写

非同期的遠隔データのシャドウイングは、一次サイトと二次サイトを大きな距離によって隔てることによって1つの災害が一次サイト及び二次サイトの両方を不良化させる確率を減少させる必要がある時、或いは一次アプリケーション・パフォーマンスのインパクトを最小にする必要がある時に使用される。一次サイト及び二次サイトの間の距離が地球を横切って延びることができる時、多数の一次サブシステムの背後の多数の二次サブシステムまでの多数のDASDボリュームにまたがる書込み更新の同期化は更にかなり複雑になる。レコード書込み更新は、二次記憶装置サブシステムにおけるシャドウイングのために一次記憶制御装置から一次データ・ムーバを介して二次データ・ムーバまで配送可能であるが、それらの間で送られる制御データの量は最小にされなければならない。それは、幾つかの記憶制御装置の背後における多数のDASDボリュームにまたがって一次システムで生じるように、幾つかの記憶制御装置にまたがって二次システムにおいて正確な順序のレコード書込み更新を再構成することが依然としてできることを必要とする。

【0025】(b) 同期的遠隔複写

災害回復のための同期的実時間遠隔複写は、複写されたDASDボリュームがセットを形成することを必要とする。そのようなセットの形成は、更に、各セットを構成するボリューム(VOLSER)及び一次サイトの等価物を識別するための十分な量のシステム情報が二次サイトに与えられることを必要とする。重要なこととして、二次サイトは一次サイトと共に「二重ベア」を形成し、二次サイトは、1つ又は複数個のボリュームがそのセットと同期していない時、即ち、「障害ある二重」が生じた時を認識しなければならない。接続障害は、非同期的遠隔複写における接続障害よりも同期的遠隔複写における接続障害の方がずっと可視的である。それは、代わりのパスが再試行される間、一次DASD I/Oが遅れるためである。一次サイトは、二次サイトに対する更新が待ち行列化される時、その一次サイトが継続することを可能にするために複写を中止又は抑止することができる。二次サイトを示すために一次サイトがそのような更新をマークすることは同期しない。二次サイトを一次サイトとの同期外れにする例外条件を認識することは、二次サイトが災害回復のためにいつでも使用可能であるためには必要である。エラー状態及び回復アクションは二次サイトを一次サイトと不整合にしてはならない。

【0026】しかし、二次DASDが存在し且つアクセス可能である時に二次サイトと一次サイトとの間の接続を維持することはデータ内容の同期状態を保証するものではない。二次サイトは、多くの理由で一次サイトとの同期状態を緩めることが可能である。二重ベアが形成される時、二次サイトは、先ず、同期外れとなり、そして

12

初期データ複写が完了する時に同期に達する。一次サイトは、それが更新されたデータを二次サイトに書き込むことができない場合に二重ベアを解くことが可能であり、その場合、一次サイトは、更新アプリケーションが継続するように、抑止二重ベア状態の下で一次DASDに更新を書き込む。従って、一次サイトは露出状態で、即ち、二重ベアが回復するまで現在の災害保護複写なしで実行しようとする。二重ベアを回復する時、二次サイトは直ちには同期しない。未決の更新を適用した後、二次サイトは同期に戻る。一次サイトは、そのボリュームに対する抑止コマンドを一次DASDに発生することによって二次サイトに同期を失わせることもできる。二次サイトは、抑止コマンドが終了し、二重ベアが再設定され、未決の更新が複写された後に一次サイトと再同期する。オンライン保守は同期化を失わせることもできる。

【0027】二次ボリュームが一次ボリュームと同期していない時、二次ボリュームは二次システム回復及び一次アプリケーションのリザンプションに対して使用可能ではない。二次サイトにおける同期外れのボリュームは、そのように識別されなければならない。二次サイト回復引継手順は、アプリケーション・アクセスを否定する(ボリューム・オフラインを強制するか、或いはそれらのVOLSERを変更する)ための同期外れボリュームを識別する必要がある。二次サイトは、一次サイトのホストがアクセス不能である任意の時間に一次サイトを回復させるよう要求可能であり、従って、二次サイトは、すべてのボリュームの同期状態に関するすべての関連情報を必要とする。二次記憶制御装置である二次記憶装置サブシステムは、一次サイトに一次サイト遭遇の例外による同期化をブレイクさせるすべての条件を決定することはできない。例えば、二次サイトが知らない一次I/Oパス又はリンクの障害のために一次サイトが二次ピアをアクセスすることができない場合、一次サイトは二重ベアをブレイクすることが可能である。この場合、二重ベアがブレイクされることを一次サイトが表す時、二次サイトは同期状態を示す。

【0028】同期外れの二重ベア・ボリュームが存在することを、外部コミュニケーションが二次サイトに知らせることがある。これは、ユーザ・システム管理機能を使用することによって実現可能である。一次I/Oオペレーションはチャンネル終了/装置終了/ユニット・チェック(CE/DE/UC)状態でもって終了し、センス・データはそのエラーの性質を表す。この形式のI/O構成によって、エラー回復プログラム(ERP)はそのエラーを処理し、I/Oが終了したことを一次アプリケーションに知らせる前に適当メッセージを二次プロセッサに送る。そこで、ユーザは、ERPが二重ベア・メッセージを抑止したことを認識してその情報を二次ロケーションに確保するように応答可能である。二次サイトが一次サイトの場所で動作的になることに依存する時、始

13
動手順は二次DASDオンラインを二次ホストに与える。その場合、二次DASDサブシステムに記憶された同期ステータスは、同期外れボリュームがアプリケーション割当てのためのオンラインにされないことを保証するために検索される。すべてのERP抑止二重ベア・メッセージとマージされたこの同期ステータスは二次同期外れボリュームの完全な画像を与える。

【0029】図1を参照すると、一次サイト14及び二次サイト15を有する災害回復システム10が示される。この場合、二次サイト15は、例えば、一次サイト14から20キロメートル離れて位置している。一次サイト14は一次ホスト・プロセッサ1を含み、それはそこで走るアプリケーション及びシステムI/O及びエラー回復プログラム2（以後、I/O&ERP2と呼ぶ）を有する。一次プロセッサ1は、例えば、DFSMS/MVSオペレーティング・ソフトウェアで走るIBMエンタープライズ・システム/9000（ES/9000）プロセッサでよく、そこで走る幾つかのアプリケーション・プログラムを持つものでよい。一次記憶制御装置3、例えば、IBM3990モデル6記憶制御機構は、チャンネル12を介して一次プロセッサ12に接続される。その分野では知られているように、幾つかのそのような記憶制御装置3が一次プロセッサ1に接続可能であり、或いは、幾つかの一次プロセッサ1が一次記憶制御装置3に接続可能である。一次DASD4、例えば、IBM3390DASDは一次記憶制御装置3に接続される。幾つかの一次DASD4が一次記憶制御装置3に接続可能である。一次記憶制御装置3及びそれに接続された一次DASD4は一次サブ記憶システムを形成する。更に、一次記憶制御装置3及び一次DASD4は単一の統合ユニットであってもよい。

【0030】二次サイト15は二次ホスト・プロセッサ5、例えば、IBM ES/9000を含み、それはチャンネル13を介して二次記憶制御装置6、例えば、IBM3990モデル6に接続される。更に、DASD7がその記憶制御装置6に接続される。一次プロセッサ1は少なくとも1つのホスト間コミュニケーション・リンク11、例えば、チャンネル・リンク又は電話T1/T3ライン・リンク等によって二次プロセッサ5に接続される。一次プロセッサ1は、例えば、多数のエンタープライズ・システム接続（ESCOM）リンク9による二次記憶制御装置6との直接接続を持つようにしてもよい。その結果、I/O&ERP2は、必要に応じて二次記憶制御装置6とコミュニケーションすることができ、一次記憶制御装置3は多数の同格通信リンク8、例えば、多数のESCOMリンクを介して二次記憶制御装置6とコミュニケーションする。

【0031】書込みI/Oオペレーションが一次プロセッサ1において走るアプリケーション・プログラムによって実行される時、そのI/Oオペレーションが成功裏

14
に完了したことを表すハードウェア・ステータスのチャネル終了/装置終了（CE/DE）が与えられる。一次プロセッサ1のオペレーティング・システム・ソフトウェアは、I/Oオペレーションの成功した完了時にアプリケーションに書込みI/Oオペレーションの成功をマークし、従って、アプリケーション・プログラムが、成功裏に完了した第1の又は前の書込みI/Oオペレーションに従属した次の書込みI/Oオペレーションに継続することを許容する。一方、その書込みI/Oオペレーションが不成功であった場合、チャネル終了/装置終了/ユニット・チェック（以後、CE/DE/UCと呼ぶ）のI/Oステータスが一次プロセッサ1のオペレーティング・システム・ソフトウェアに与えられる。ユニット・チェックを与えると、I/O&ERP2は、障害のあったI/Oオペレーションの性質に関する特殊セン

10
ス情報を一次記憶制御装置3から得る制御を行う。或ボリュームに特有のエラーが生じた場合、そのエラーに関連した特有のステータスがI/O&ERP2に与えられる。しかる後、I/O&ERP2は、一次記憶制御装置3及び二次記憶制御装置6との間の、或いは、最悪の場合には、一次プロセッサ1及び二次プロセッサ5との間のデータ整合性を維持するために新しい同格通信同期エラー回復を行うことができる。

【0032】図2を参照すると、エラー回復手順が示される。図2において、ステップ201は、一次プロセッサ1において走るアプリケーション・プログラムがデータ更新を一次記憶制御装置3に送ることを含む。ステップ203において、そのデータ更新が一次DASD4に書き込まれ、そして、そのデータ更新は二次記憶制御装置6にシャドウされる。ステップ205において、一次サイト及び二次サイトが同期しているかどうかを決定するために、二重ベア・ステータスがチェックされる。その二重ベア・ステータスが同期状態にある場合、一次プロセッサ1においてそこで走るアプリケーション・プログラムを介して処理が継続する間、ステップ207において、データ更新が二次DASD7に書き込まれる。

【0033】二重ベアが「障害」状態にある場合、ステップ209において、一次記憶制御装置3は、その二重ベアが抑止したこと或いは障害を生じたことを一次プロセッサ1に知らせる。その二重ベアは、コミュニケーション・リンク8を介した一次記憶制御装置3及び二次記憶制御装置6の間のコミュニケーション障害のために「障害」となることがある。それとは別に、二重ベアは一次サブシステム或いは二次サブシステムにおけるエラーのために「障害」となることがある。その障害がコミュニケーション・リンク8にある場合、一次記憶制御装置3は、二次記憶制御装置6に直接にその障害をコミュニケーションすることができない。ステップ211において、一次記憶制御装置3はI/OステータスCE/DE/UCを一次プロセッサ1に戻す。I/O&ERP2は

(9)

特開平8-305500

15

アプリケーション・プログラムを静止させ、従って、書込みI/Oオペレーションをリクエストするアプリケーションに制御を戻す前に、エラー回復及びデータ整合性のために、ステップ213において、一次プロセッサ1の制御を行う。

【0034】図3を参照すると、記憶制御装置325、例えば、IBM3990記憶制御機構は、例えば、データ機能記憶管理サブシステム/多重仮想システム(DFSMS/MVS)を走らせるIBMシステム/370或いはIBMエンタープライズ・システム/9000(ES/9000)プロセッサのようなホスト・プロセッサ310を含むデータ処理システムに接続されるものとして更に詳細に示される。記憶制御装置325は、更に、IBM3380又は3390DASDのような直接アクセス記憶装置(DASD)375に接続される。記憶サブシステムは記憶制御装置325及びDASD375によって形成される。その記憶サブシステムは、コミュニケーション・リンク321を介してホスト・プロセッサ310に接続される。そのコミュニケーション・リンク321は、ホスト・プロセッサ310のチャンネル320に及び記憶制御装置325のポートA-D、E-Hに接続する。コミュニケーション・リンク321は、並列又は直列リンク、例えば、エンタープライズ・システム接続(ESCOM)直列ファイバ光学リンクであってもよい。

【0035】記憶制御装置325は二重クラスタ360、361を含む。その二重クラスタ360、361は別々の電源(図示されていない)を有し、更に、コミュニケーション・インターフェースを与えるためのポートA-D、E-H330を含む。不揮発性記憶装置(NVM)370及びキャッシュ445の両方が一時的データ記憶装置に与えられ、クラスタ360、361の両方にアクセス可能である。記憶装置バス0-3340がDASD375への必要なバスを与える。重要製品データがVPD395及び396に維持される。記憶制御装置325と同様の記憶制御装置が米国特許第5,051,887号に開示されている。

【0036】図4は記憶制御装置の記憶装置バス401を更に詳細に示す。前に図3に示されたように、記憶制御装置は4つの記憶装置バスを有し、各記憶装置バスは他の3つと同じである。従って、1つの記憶装置バスだけを詳細に説明することにする。記憶装置バス401は上部のチャンネル・ポート430によって8×2スイッチ402に接続され、下部の装置ポート432によって複数のDASDに接続される。記憶装置バス401は、そのバス内で生じるすべてのオペレーションを制御するマイクロプロセッサ410を含む。マイクロプロセッサ410は、ホスト・プロセッサから受け取ったチャンネル・コマンドを翻訳することができ、接続されたDASDを制御することもできる。マイクロプロセッサ410

16

は、外部サポート機構を通して制御メモリ又は制御記憶装置(図示されていない)にロードされたマイクロ命令を実行する。

【0037】図4には、共用制御アレイ(SCA)434も示される。SCA434はその記憶制御装置の4つの記憶装置バスすべてによって共用される情報を含む。記憶装置バス401における各マイクロプロセッサ410は共用の情報を得るためにSCA434をアクセスする。一般的な共用の情報は、4つの記憶装置バスすべてのマイクロプロセッサにより使用される外部レジスタ、装置ステータス、及びチャンネル再接続データを含む。

【0038】記憶装置バス401はポート・アダプタ(PA)412を含み、そのポート・アダプタはキャッシュ420、不揮発性記憶装置(NVM)422、及び自動データ転送(ADT)バッファの間でデータを転送するためのデータ・バス及び制御ラインを与える。そのADTバッファはADT回路414及び速度変更バッファ416より成る。速度変更バッファ416は、そのDASDのデータ転送速度及びチャンネルに対するホスト・プロセッサのデータ転送速度の間の差を補償する。一般に、データ処理システムでは、チャンネルと記憶制御装置との間のデータ転送速度、又はチャンネル転送速度は、DASDと記憶制御装置との間のデータ転送速度又はDASD転送速度よりもずっと高い。

【0039】ポート・アダプタ412は上位キャッシュ・ポート424及び下位キャッシュ・ポート426を使用してキャッシュ420、NVS422、及びADTバッファ(414、416より成る)の間のデータ・バスを与える。これらの2つのポートは、キャッシュ420に関連する2つの同時転送を可能にする。例えば、下位キャッシュ・ポート426を使用してデータがDASDからキャッシュ420に転送されるのと同時に、上位キャッシュ・ポート424を使用してデータがキャッシュ420からチャンネルに転送可能である。データ転送はマイクロプロセッサ410によって開始され、そしてそれが一旦開始されると、完了するまでマイクロプロセッサの介入なしにADT回路414によって制御される。

【0040】記憶装置バス401は、直接DASDオペレーション時に、又はキャッシュ・オペレーション時に、又は高速書込みオペレーション時に、ホスト・プロセッサから複数のDASDの1つへのデータ・レコードの転送を指示する。直接DASDオペレーションは、データの一時記憶のためにキャッシュ又はNVSを使用することなく、ホスト・プロセッサと複数のDASDの1つとの間のデータの転送を含む。この場合、記憶装置バス401はADTバッファ414、416を使用して、DASDへ転送するためのデータを一時的に記憶する。

【0041】キャッシュ・オペレーション時に、記憶装置バス401はデータをキャッシュ・メモリ420に記

(10)

特開平8-305500

17

憶し且つそのデータをDASDにブランチする。この場合、そのデータは上部のチャネル・ポート430を使用してADTバッファ414、416転送される。そこで、データは、上位キャッシュ・ポート424を使用してADTバッファ414、416からキャッシュ・メモリ420に転送され、下部の装置ポート432を使用してDASDに転送される。データは、それがDASDにブランチされた後、或時間的期間の間キャッシュ・メモリ420に残っている。ホスト・プロセッサがデータを、それが更新される前に読み取るようリクエストする場合、記憶装置バス401はそのデータをキャッシュ420から読み取るよう指示することができ、それによってデータ処理システムのパフォーマンスを向上させることができる。

【0042】高速の書込みオペレーション時に、記憶装置バス401は最初にデータをキャッシュ420及びNV S 4 2 2に記憶する。そのデータは、その後、NV S 4 2 2からDASDにデステージされる。この高速書込みの場合、データは上部のチャネル・ポート430を使用してADTバッファ414、416に転送される。しかる後、データはADTバッファ414、416から上位キャッシュ・ポート424を使用してキャッシュ420へ及び下位キャッシュ・ポート426を使用してNV S 4 2 2へ転送される。キャッシュ・オペレーションの時のように、ホスト・プロセッサがデータを、それが更新される前に読み取るようリクエストされる場合、記憶装置バス401はデータをキャッシュ420から読み取るよう指示することができ、それによって、データ処理システムのパフォーマンスを向上させることができる。

【0043】図5を参照すると、二重複写災害回復機能及び遠隔二重複写災害回復機能の両方を実施できるデータ処理システムが示される。ホスト・プロセッサ501、例えば、IBM ES/9000は2つの記憶サブシステム502、503にコミュニケーションする。記憶制御装置510、例えば、IBM3990モデル3は、IBM3990のようなDASD512、514、516、522、524及び526に接続され、ホスト・プロセッサ501と同じ一次サイトに置かれた1つの記憶サブシステム502を構成する。この記憶サブシステム502は一次DASD512、514、及び516に書かれたデータを、それぞれ二次DASD522、524、及び526上にバックアップする二重複写オペレーションを遂行する。DASD512、522は1つの二重ペアを形成する。同様に、2つの更なる二重ペアがDASD514、524及びDASD516、526によって形成される。

【0044】ホスト・プロセッサ501において走るアプリケーション・プログラムが記憶サブシステム502に書き込まれるべきレコードを発生する時、ホスト・プロセッサ501は、先ず、そのレコードを記憶サブシ

18

テム502の記憶制御装置510に転送する。記憶制御装置510は、そのレコードを受け取り、転送の成功を信号するチャネル終了(CE)をホスト・プロセッサに発生する。そこで、記憶制御装置510は、そのレコードを一次DASD512へ転送し、一次DASD512へのレコードの書込みの成功を信号する装置終了(DE)をホスト・プロセッサ501に発生する。ホスト・プロセッサ501は、今や、その後のレコードを一次DASD512、514、516の1つに書き込むか、或いは前に書き込まれたデータを一次DASD512、514、516から読み取るという記憶サブシステム502に対する次のオペレーションを遂行することができる。そこで、その記憶制御装置は、一次DASD512に書き込まれたレコードの複写を二次DASD522に転送し、そのレコードのバックアップ・バージョンを与える。

【0045】ホスト・プロセッサ501は、記憶制御装置530、例えば、IBM3990モデル6、及び幾つかのDASD532、534、536、例えば、IBM3390或いはIBM RAMACより成る他の記憶サブシステム503に接続される。この記憶サブシステム503は災害回復システムのための一次サイトを与える。DASD532、534、536は二重ペアの一次DASDとして働く。一次記憶サブシステム503は、災害回復システムを完成するための第2の遠隔サイトにおける二次記憶サブシステム504に接続される。一次記憶サブシステム503及び二次記憶サブシステム504の間の接続550は直接コミュニケーション・リンク、例えば、エンタープライズ・システム接続(ES/COM)リンクを通して行われる。二次記憶サブシステム504は、IBM3390或いはIBM RAMACのような幾つかのDASD542、544、546に接続された記憶制御装置540、例えば、IBM3990モデル6より成る。二次記憶サブシステム504におけるDASD542、544、546は遠隔二重複写オペレーションに対する二次DASDとして働く。図5における災害回復システムには、3つの二重ペアが示される。DASD532、542は第1二重ペアを形成し、DASD534、544は第2二重ペアを形成し、DASD536、546は第3二重ペアを形成する。従って、例えば、ホスト・プロセッサ501は一次記憶サブシステム503の記憶制御装置530を通して一次DASD532にレコードを書き込む。このレコードの複写は、その後、一次記憶制御装置530によって二次記憶制御装置540に転送され、その二重ペアの二次DASD542に記憶される。

【0046】ホスト・プロセッサにおいて走るアプリケーションは記憶サブシステムにI/Oオペレーションを発生する。I/Oオペレーションの例は、DASDからの読取り、DASDへの書込み、及びデータの転送を必

(11)

特開平8-305500

19

要としないDASDに対する他のコマンドである。アプリケーションは、I/Oに無関係の他のI/Oオペレーションを条件としないI/Oオペレーション、又はI/Oに従属した他のI/Oオペレーションを条件とするI/Oオペレーションを発生することができる。従属したI/Oの例として、アプリケーションは、一次DASD 532にレコードを書き込むことによって第1 I/Oを発生し、しかる後、レコードを指示するインデックスを他の一次DASDに書き込むことによって第2 I/Oを発生してもよい。第2 I/Oは従属I/Oである。

【0047】従属I/Oは、データ処理システムが多数のサブシステムに跨って多数の二重ベアを抑止しようとする時に問題を生じることがある。その問題は、第2 I/Oに対する二重ベア534、544が抑止される前にレコードを指示するインデックスが二次DASD 544に複写されるが、そのレコードが二次DASD 542に複写される前に第1 I/Oに対する二重ベア532、542が抑止される場合に生じる。この例では、二次DASDにおけるデータはシーケンス整合した順序で同期されない。レコードに対するインデックスは二次DASD 544に複写されているが、そのレコードは二次DASD 542に複写されてない。従って、多数の記憶サブシステムに跨って多数の二重ベアを抑止することは、第1 I/Oがその二重ベアの二次DASDに複写されない場合、第2の従属したI/Oがその二重ベアの二次DASDに複写できないことを必要とする。

【0048】図6を参照すると、二次装置におけるデータがシーケンス整合した順序で同期したままであるように多数の二重ベアを抑止するための方法を説明する流れ図が示される。ステップ610において、ホスト・プロセッサは特殊な二重ベアに対する静止二重ベア・コマンドを記憶制御装置に発生する。記憶制御装置はそのコマンドを受け、その二重ベアを静止させる。二重ベアを静止させることは、進行中の何れの現在のI/Oオペレーションも完了することを可能にし、指定された二重ベアに対する如何なる将来のI/Oオペレーションも記憶制御装置又はホスト・プロセッサにおいて待ち行列化されないようにする。ステップ620は、更なる二重ベアが静止される必要があるかどうかを決定する。それが肯定される場合、ステップ610が更なる二重ベアの各々に対して繰り返される。それが否定される場合、ステップ630は、すべての二重ベアが静止させられそして二重ベアの二次DASDにおけるデータがシーケンス整合した順序で同期化されることを表す。二次DASDにおけるシーケンス整合性を得る手順の詳細が図7に示される。

【0049】ステップ640において、ホスト・プロセッサは指定された二重ベアに対する抑止二重ベア・コマンドを記憶制御装置に発生する。記憶制御装置はそのコマンドを受け取り、二重ベアを抑止する。その二重ベア

20

を抑止することは静止を解放し、一次システム・アプリケーションの読取り及び書込みが二重ベアの一次DASDに再開することを可能にする。一方、抑止された状態は、これらの変化を二重ベアの二次DASDに複写することを妨げる。ホスト・プロセッサから一次DASDへのその後の書込みオペレーションは記憶制御装置によって二次DASDに複写されない。代わりに、記憶制御装置は、ホスト・プロセッサからのコマンドに応答して二重ベアを終了させるか、或いはその後の書込みオペレーションの物理的DASDアドレス、即ち、特定の一次DASDにおける物理的ロケーションをレコードすることができる。二重ベアが終了する場合、その二重ベアが抑止された後に一次DASDに書き込まれたレコードは、その二重ベアがその後に再設定される時には二次DASDに複写されないであろう。しかし、二重ベアがレコーディングの変更を抑止される場合、その二重ベアが抑止された後に一次DASDに書き込まれたレコードは、その二重ベアがその後に再設定される時、二次DASDに複写されるであろう。

【0050】ステップ650は、更なる二重ベアが抑止される必要があるかどうかを決定する。それが肯定される場合、ステップ640が更なる二重ベアの各々に対して繰り返される。それが否定される場合、ステップ660は、すべての二重ベアが抑止されそしてその二重ベアの二次DASDにおけるデータがシーケンス整合順序で同期化されることを表し、ホスト・プロセッサにおけるアプリケーションが一次DASDを使用して走ろうとしていることを表す。

【0051】図7は、二重ベアを静止させるために取られるステップを表す流れ図を示す。ステップ710は、ホスト・プロセッサからの静止二重ベア・コマンドが記憶制御装置において受信されたかどうかを決定する。そのようなコマンドが受信されなかった場合、静止二重ベア・プロセスは終了する。それが肯定された場合、記憶制御装置は静止二重ベア・コマンドに応答してその指定された二重ベアの一次DASDに特有な長いビジー・フラッグをセットする。ステップ720は、記憶制御装置が、静止コマンドの前に開始し且つ現在進行中の二重ベアに対する何れの書込みオペレーションを完了させることを表す。ステップ730において、長いビジー・フラッグは、ホスト・プロセッサによる二重ベアへのその後のI/Oリクエストに回答して、ホスト・プロセッサへの長いビジー信号を生じるように記憶制御装置に指示する。従って、その長いビジー信号は、長いビジー・オペレーションが終了するまで、ホスト・プロセッサからのその後のI/Oオペレーションをホスト・プロセッサにおいて待ち行列化させる。この待ち行列化は記憶制御装置においても生じ得るが、ホスト・プロセッサはよりよい待ち行列プラットフォームを与える。

【0052】図6及び図7に示された方法は、図5の構

(12)

特開平8-305500

21

成を使用して説明可能である。ホスト・プロセッサ501は、記憶サブシステム503の一次DASD534に第1 I/Oを書き込み、しかる後、記憶サブシステム503の一次DASD532に第2 I/O、即ち、従属 I/Oを書き込む。そこで、ホスト・プロセッサ501は、記憶サブシステム502、503、504における6つの二重ベアすべてに対して連続した静止二重ベア・コマンドを発生する。二重ベア534、544に対するその静止コマンドは、二重ベア532、542に対する静止コマンドの前に発生される必要はない。

【0053】二重ベア532、542が静止させる時、二重ベア534、544に対する第1 I/Oは次の3つの可能な状態のうちのどれかにある。即ち、

(1) その I/Oは、一次DASD534及び二次DASD544の両方に対して完了可能である。

(2) その I/Oは、記憶制御装置510によって長いビジーに保持可能であり、従って、一次DASD534或いは二次DASD544に未だ転送可能ではない。

(3) その I/Oは、一次DASD534に対して進行中であり、記憶制御装置530が長いビジーでもってその後の I/Oを締め出す前に二次DASD544に対して完了可能である。

【0054】第1 I/Oに対して状態2が生じた場合、二重ベア532、542に対する第2 I/Oは、それが従属 I/Oであり且つ第1 I/Oの完了を条件とされないで発生されなかったであろう。従って、第1 I/O及び第2 I/O共、二重ベアを静止させる前に転送されないであろうし、二次DASD534及び542はシーケンス整合性を維持するであろう。

【0055】第1 I/Oに対して状態1又は3の何れかが生じた場合、第2 I/Oは第1 I/Oと同様に次の3つの可能な状態のうちのどれかにある。即ち、

(A) その I/Oは、一次DASD532及び二次DASD542の両方に対して完了可能である。

(B) その I/Oは、記憶制御装置530によって長いビジーに保持可能であり、従って、一次DASD532或いは二次DASD542には未だ転送可能ではない。

(C) その I/Oは、一次DASD532に対して進行中であり、記憶制御装置530が長いビジーでもってその後の I/Oを締め出す前に二次DASD542に対して完了可能である。

従属 I/Oに対して状態Bが生じた場合、その従属 I/Oはその静止コマンド前に二重ベア532、542に転送されないが、第1 I/Oはその静止コマンドの前に二重ベア534、544に対して完了する。しかし、これは二次DASD544及び542においてシーケンス整合性を維持する。

【0056】第1 I/Oに対する状態1又は3と関連して状態A又はCが従属 I/Oに対して生じた場合、シーケンス整合性が二次DASD544及び542において

22

維持される。何れの組合せにおいても、第1 I/O及び従属 I/Oはそれぞれの二次DASD544及び542に対して完了する。その状況は、従属 I/Oがその二次DASD542に複写され且つ第1 I/Oがその二次DASD544に複写されない場合には生じない。従って、二次DASDは同期化され、一次DASDが更新されたシーケンスと整合した順序で維持される。

【0057】この例は同期的遠隔複写構成における二重ベアを扱ったけれども、その方法はサブシステム501のような二重複写構成における二重ベアにも、及び遠隔複写構成或いは二重複写構成における多数の記憶制御装置にまたがる二重ベアにも適用する。

【0058】本発明をその好適な実施例に関連して詳しく示し且つ説明したけれども、本発明の精神及び技術範囲を逸脱することなく、形式及び詳細における種々の変更を行うことが可能であることは当業者には明らかなことであろう。例えば、レコード更新のフォーマットは決定的なものではなく、そのようなフォーマットはCKD、ECKD、固定ブロックアーキテクチャ(FBA)等であってもよい。更に、記憶装置はDASD装置に限定されることを意味するものではない。

【0059】まとめとして、本発明の構成に関して以下の事項を開示する。

(1) レコード及びレコード更新を書き込むこと及びバックアップの目的で前記レコード及びレコード更新を複写することができるデータ記憶システムにおいて、前記データ記憶システムはレコード及びレコード更新を発生するアプリケーションを走らせるホスト・プロセッサと、一次データ記憶装置及び二次データ記憶装置を有する第1二重ベアと、一次データ記憶装置及び二次データ記憶装置を有する第2二重ベアと、前記ホスト・プロセッサと前記第1二重ベア及び第2二重ベアの前記一次データ記憶装置との間に接続された記憶制御装置とを含み、レコード及びレコード更新の複写が進行中である時に第1及び第2二重ベアを抑止するための方法にして、前記アプリケーションから前記第1二重ベアの一次データ記憶装置への将来のレコード及びレコード更新の書き込みを禁止するために前記ホスト・プロセッサからの初期静止コマンドに応答して前記記憶制御装置によって前記第1二重ベアを静止させるステップと、前記アプリケーションから前記第2二重ベアの一次データ記憶装置への将来のレコード及びレコード更新の書き込みを禁止するために及びシーケンス整合した順序における前記第1及び第2二重ベアの前記二次データ記憶装置に複写されたレコード及びレコード更新を同期化するために前記ホスト・プロセッサからのその後の静止コマンドに応答して前記記憶制御装置によって前記第1二重ベアを静止させるステップと、前記アプリケーションから前記第1及び第2二重ベアの一次データ記憶装置に送られたレコード及びレコード更新を前記第1及び第2二重ベアの二次デー

23

タ記憶装置に前記記憶制御装置によって複写することを禁止するために、前記記憶制御装置が前記ホスト・プロセッサから抑止コマンドを受けることによって前記第1及び第2二重ベアを抑止するステップと、前記第1及び第2二重ベアが抑止されること及びその後のレコード及びレコード更新が前記第1及び第2二重ベアの前記二次データ記憶装置に前記レコードを複写することなく前記第1及び第2二重ベアを前記一次データ記憶装置に書き込み可能であることを前記記憶制御装置によって前記アプリケーションに信号するステップと、を含む方法。

(2) 前記静止させるステップは前記アプリケーションから前記二重ベアの一次データ記憶装置へのレコード及びレコード更新のその後の書き込みを禁止するための長いビジー信号を前記記憶制御装置から前記ホスト・プロセッサに発生するステップを含むことを特徴とする上記(1)に記載の方法。

(3) 前記記憶制御装置は、前記二重ベアが再設定される場合、前記ホスト・プロセッサから前記一次データ記憶装置に転送されたその後のレコード及びレコード更新が前記二次データ記憶装置へのその後の複写のためにマークされないように前記二重ベアを終了させることを特徴とする上記(1)に記載の方法。

(4) 前記記憶制御装置は、前記二重ベアが再設定される場合、前記レコード及びレコード更新が前記二次データ記憶装置に複写されるように、前記二重ベアを抑止の後に前記ホスト・プロセッサから前記一次データ記憶装置に転送されたレコード及びレコード更新をマークすることを特徴とする上記(1)に記載の方法。

(5) レコード及びレコード更新を第1及び第2二重ベアに複写することができるデータ処理システムにおいて、レコード及びレコード更新の複写が進行中である時に各二重ベアを抑止するための記憶制御装置にして、前記データ記憶システムはアプリケーションを走らせるホスト・プロセッサを含み、前記第1及び第2二重ベアの各々は一次データ記憶装置及び二次データ記憶装置を有し、前記記憶制御装置はホスト・プロセッサと前記第1及び第2二重ベアの前記一次データ記憶装置との間に接続され、前記アプリケーションはレコード及びレコード更新を発生し及び静止二重ベア・コマンド及び抑止二重ベア・コマンドを発生し、前記一次データ記憶装置は前記レコード及びレコード更新を記憶し、前記二次データ記憶装置は前記レコード及びレコード更新の複写を記憶するものにおいて、前記ホスト・プロセッサと前記第1及び第2二重ベアとの間のレコード及びレコード更新を指示するための記憶装置バスであって、前記アプリケーションからの前記静止二重ベア・コマンドに応答して前記第1及び第2二重ベアを静止させ、前記レコード及びレコード更新が前記二重ベアの各々における前記二次データ記憶装置の各々に複写され且つシーケンス整合順序で同期化されるように前記アプリケーションからの前記抑止

(13)

特開平8-305500

24

二重ベア・コマンドに応答して前記第1及び第2二重ベアを抑止する記憶装置バスと、前記記憶装置バスに接続され、前記二重ベアの二次データ記憶装置に複写されるべき前記レコード及びレコード更新を記憶するためのメモリと、を含む記憶制御装置。

(6) 前記メモリはキャッシュ・メモリであることを特徴とする上記(5)に記載の記憶制御装置。

(7) レコード及びレコード更新をシーケンス整合した順序で記憶するための不揮発性記憶装置(NVS)を含み、前記レコード及びレコード更新はその後前記二重ベアの二次データ記憶装置に複写されることを特徴とする上記(5)に記載の記憶制御装置。

(8) 前記二重ベアが抑止される時、前記レコード更新を受ける前記二重ベアと関連した物理的アドレスをマークするために前記NVSにおけるビットマップを含み、前記レコード更新は前記二重ベアが抑止されている時に前記一次データ記憶装置に書き込まれ、一旦前記二重ベアが再設定されると前記二次データ記憶装置に複写されることを特徴とする上記(7)に記載の記憶制御装置。

(9) 前記記憶装置バスは、前記ホスト・プロセッサから前記二重ベアの一次データ記憶装置に現在書き込まれている前記レコード更新が完了することを可能にするために、及びその後のレコード更新が前記ホスト・プロセッサから前記二重ベアの一次データ記憶装置に書き込まれることを禁止するために、前記静止二重ベア・コマンドに応答して長いビジー信号を前記ホスト・プロセッサに発生することを特徴とする上記(5)に記載の記憶制御装置。

(10) 複数のレコード更新を発生するアプリケーションを走らせ、及び静止二重ベア・コマンド及び抑止二重ベア・コマンドを発生するホスト・プロセッサと、前記ホスト・プロセッサに接続されたチャンネルと、各々が一次データ記憶装置及び二次データ記憶装置を有する第1及び第2二重ベアであって、前記一次データ記憶装置の各々は複数のレコード更新を記憶し、前記二次データ記憶装置の各々は前記複数のレコード更新の複写を記憶するものと、前記チャンネルによって前記ホスト・プロセッサに接続され、更に、前記第1及び第2二重ベアの一次データ記憶装置の各々に接続された記憶制御装置と、を含み、前記記憶制御装置は、前記ホスト・プロセッサから転送された複数のレコード更新を最初に記憶するためのメモリと、前記チャンネルと前記一次データ記憶装置の各々との間に接続された記憶装置バスと、を含み、前記メモリは前記記憶装置バスに接続されること、及び前記記憶装置バスは前記ホスト・プロセッサと前記一次データ記憶装置の間で前記メモリを介して前記複数のレコード更新の移動を指示し、前記ホスト・プロセッサから前記静止二重ベア・コマンド及び前記抑止二重ベア・コマンドを受け取り、前記複数のレコード更新をシーケンス整合した順序で同期化するために前記複数

(14)

25

個のレコード更新の複写を前記二次データ記憶装置に転送すること、を特徴とするデータ記憶システム。

(11) 前記記憶制御装置におけるメモリはキャッシュ・メモリであることを特徴とする上記(10)に記載のデータ記憶システム。

(12) 前記記憶制御装置において前記記憶装置バスに接続された不揮発性記憶装置(NVS)を含み、前記NVSはレコード及びレコード更新をシーケンス整合した順序で記憶すること及び前記レコード及びレコード更新は前記二重ペアの二次データ記憶装置に複写されることを特徴とする上記(10)に記載のデータ記憶システム。

(13) 前記二重ペアの各々の各二次データ記憶装置は前記一次データ記憶装置の各々に対して遠隔の二次サイトに置かれること、及び前記二次サイトは前記一次サイトにおける前記記憶制御装置から前記レコード及びレコード更新を受けるための及び前記各二次データ記憶装置に接続された記憶制御装置を含むことを特徴とする上記(10)に記載のデータ記憶システム。

(14) 前記二重ペアの各々における前記各二次データ記憶装置は前記一次データ記憶装置の各々に対して遠隔の二次サイトに置かれること、及び前記二次サイトはホスト・プロセッサと、前記ホスト・プロセッサに接続されたチャンネルと、前記チャンネルによって前記ホスト・プロセッサに接続され、更に前記各二次データ記憶装置に接続された記憶制御装置と、を含むことを特徴とする上記(10)に記載のデータ記憶システム。

(15) 前記記憶制御装置における前記記憶装置バスは、前記ホスト・プロセッサから前記二重ペアの一次デ

特開平8-305500

26

ータ記憶装置に現在書き込まれている前記レコード更新が完了することを可能にするために、及びその後のレコード更新が前記ホスト・プロセッサから前記二重ペアの一次データ記憶装置に書き込まれることを禁止するために、前記静止二重ペア・コマンドに応答して長いビジー信号を前記ホスト・プロセッサに発生することを特徴とする上記(10)に記載のデータ記憶システム。

【図面の簡単な説明】

10 【図1】 同期的遠隔データ・シャドウ機能を持った災害回復システムのブロック図である。

【図2】 図1の災害回復システムに従って同期的遠隔複写を行うためのフローチャートである。

【図3】 データ処理システムにおいて接続される記憶制御装置を更に詳細に示すブロック図である。

【図4】 データ処理システムにおける記憶制御装置に接続された記憶装置バスを更に詳細に示すブロック図である。

【図5】 多数のデータ記憶装置サブシステムにまたがる多数の二重ペアを示すブロック図である。

20 【図6】 ホスト・アプリケーションから送られたすべてのデータが二次データ記憶装置において同期化されるように多数の二重ペアを抑止するための方法のフローチャートである。

【図7】 図6に示された方法の静止ステップを更に詳細に示すフローチャートである。

【符号の説明】

10 災害回復システム

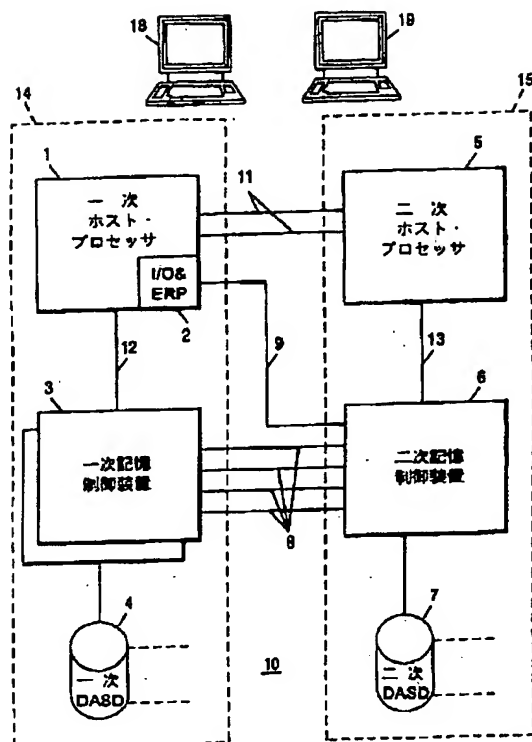
14 一次サイト

15 二次サイト

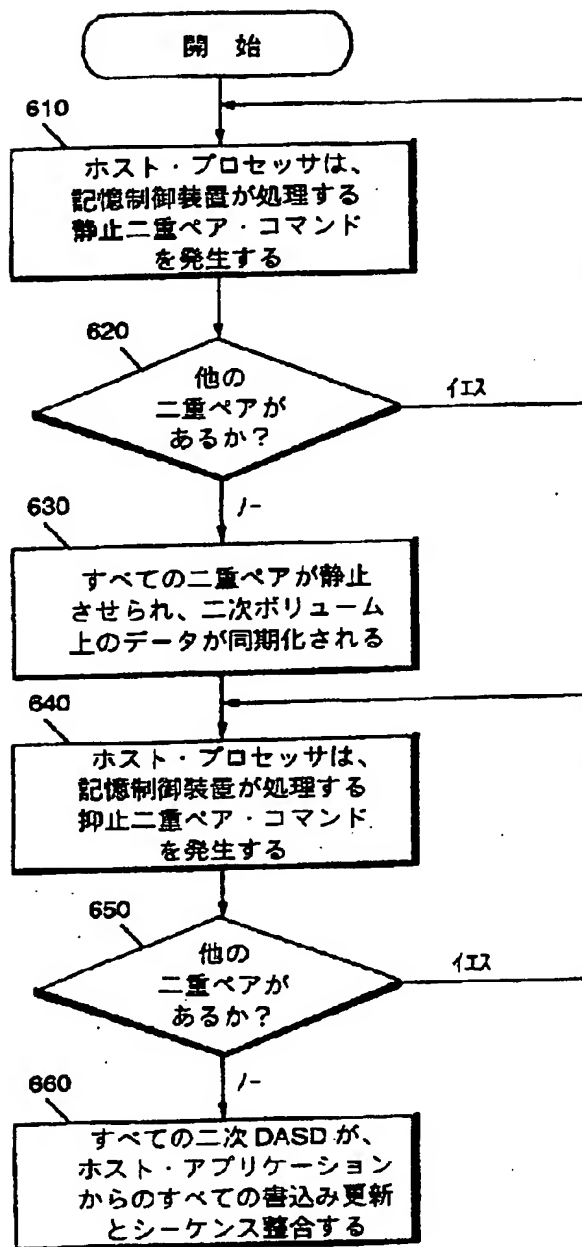
(15)

特開平 8-305500

【図 1】



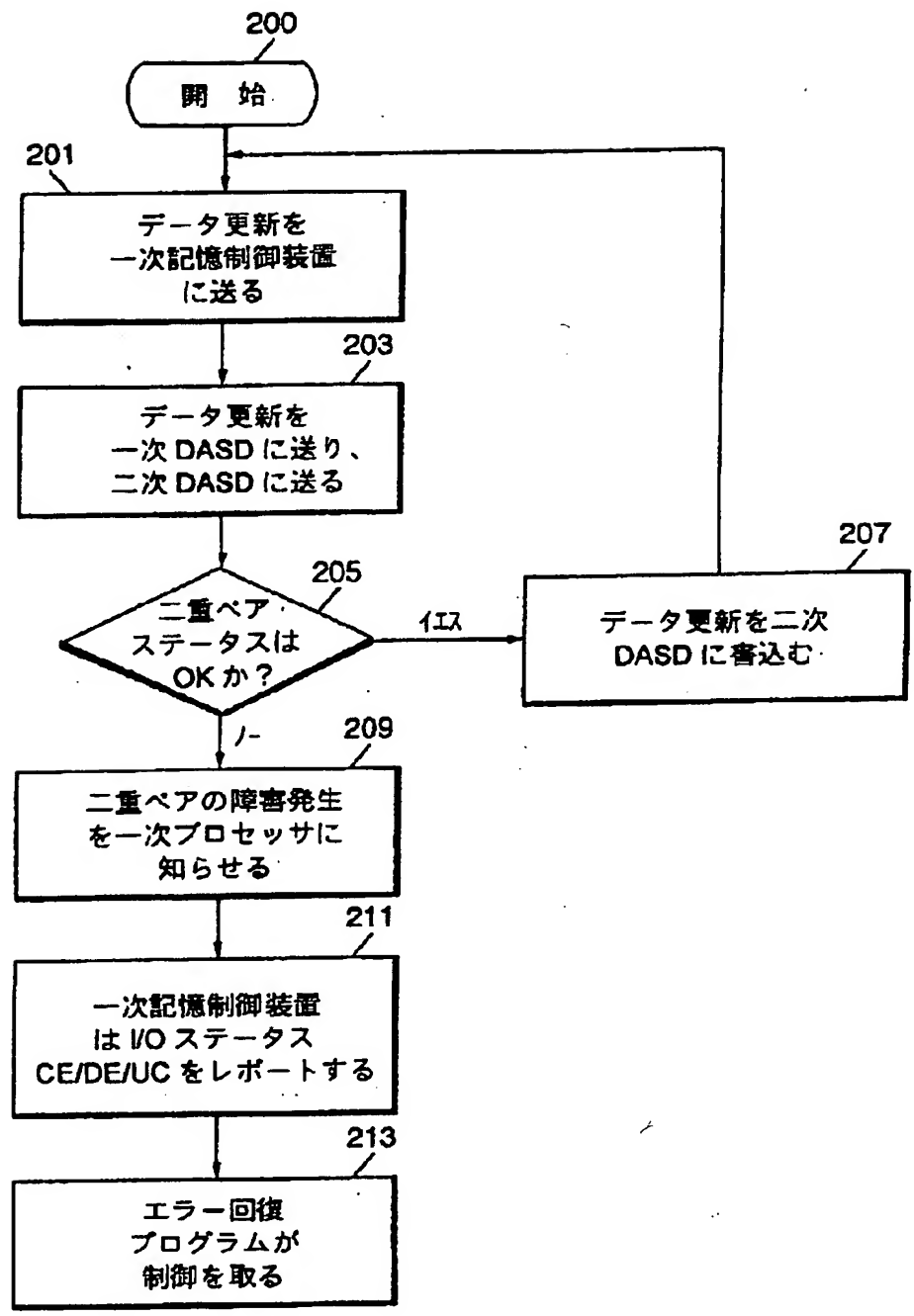
【図 6】



(16)

特開平8-305500

【図2】

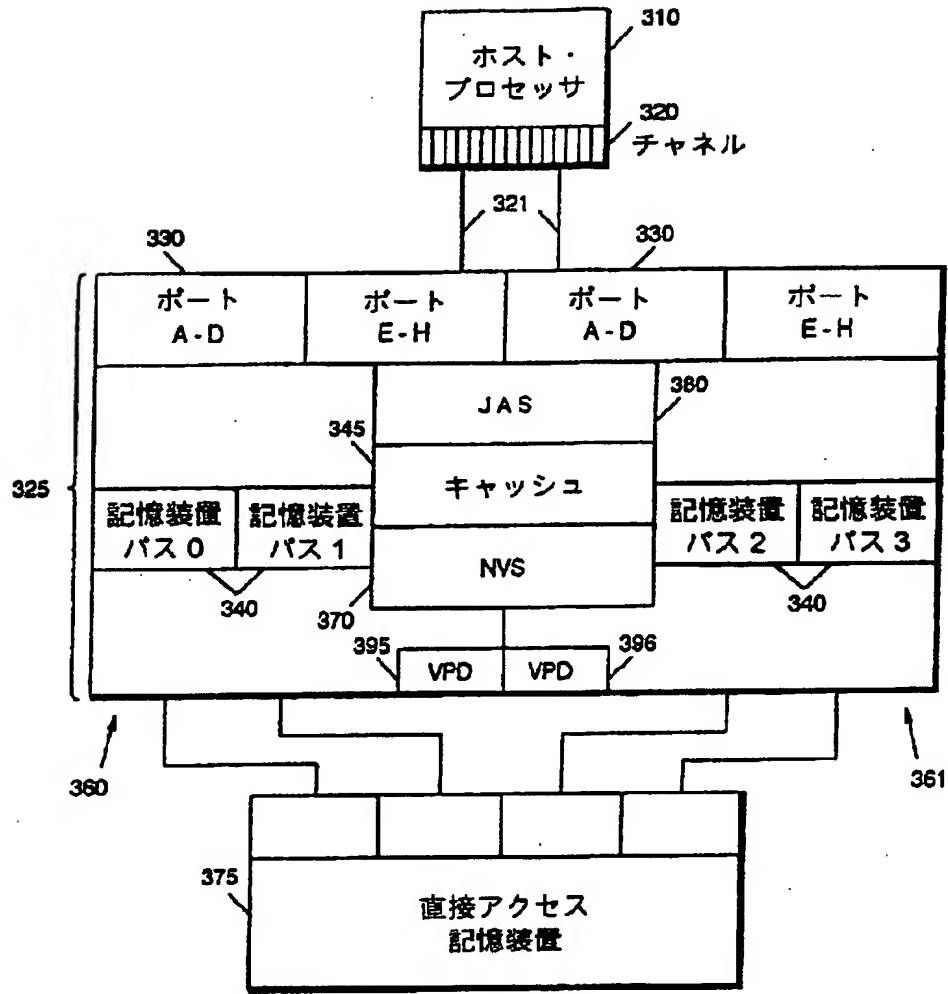


(17)

特開平 8 - 3 0 5 5 0 0

【図 3】

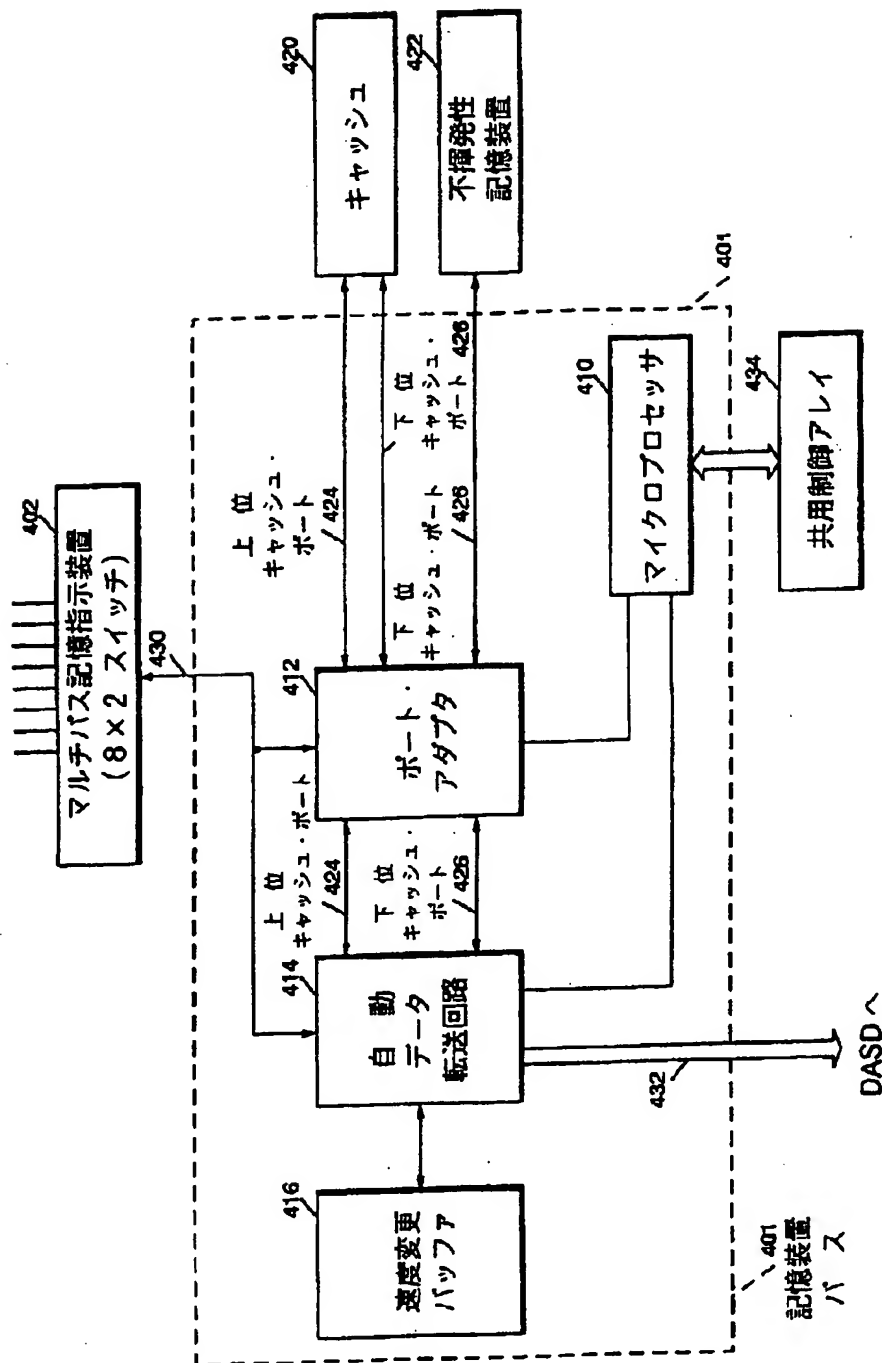
従 来 技 術



(18)

特開平8-305500

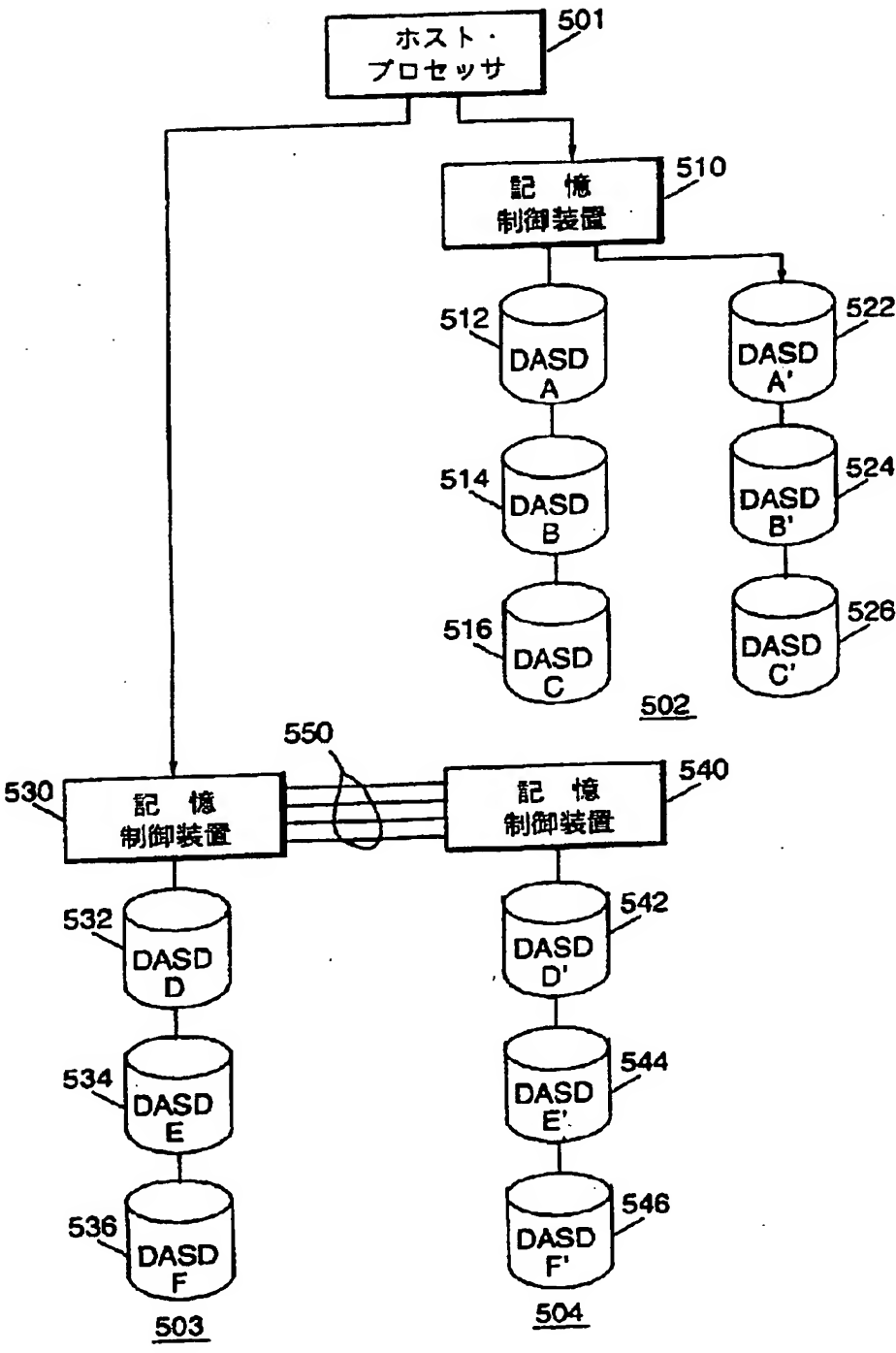
【図4】



(19)

特開平 8-305500

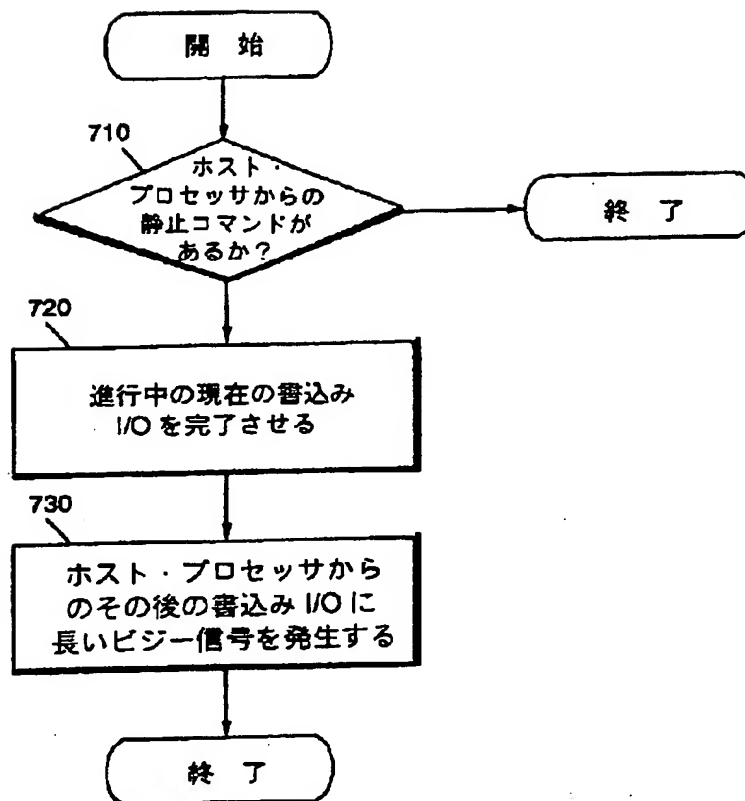
【図 5】



(20)

特開平8-305500

【図7】



フロントページの続き

(72)発明者 ロバート・フレデリック・ケーン
アメリカ合衆国アリゾナ州、ツーソン、イ
ー・コレシコ・ストリート 8338

(72)発明者 ウィリアム・フランク・ミッカ
アメリカ合衆国アリゾナ州、ツーソン、イ
ー・ラエスパルダ 3921

(72)発明者 ロバート・ウェズリー・ショムラー
アメリカ合衆国カリフォルニア州、モーガ
ン・ヒル、バイドモント・コート 17015